The principle of entropy in dialectometry. First ideas and results.

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Overview

- The principles of entropy
- Questions and problems regarding entropy
- Extending entropy
- First results: bigram analysis

The principles of entropy

- **Introduced information theory by Claude Elwood Shannon, 1948**
- **Entropy describes *how much randomness* is in an amount of data**
- **Proper Entropy H_p:**

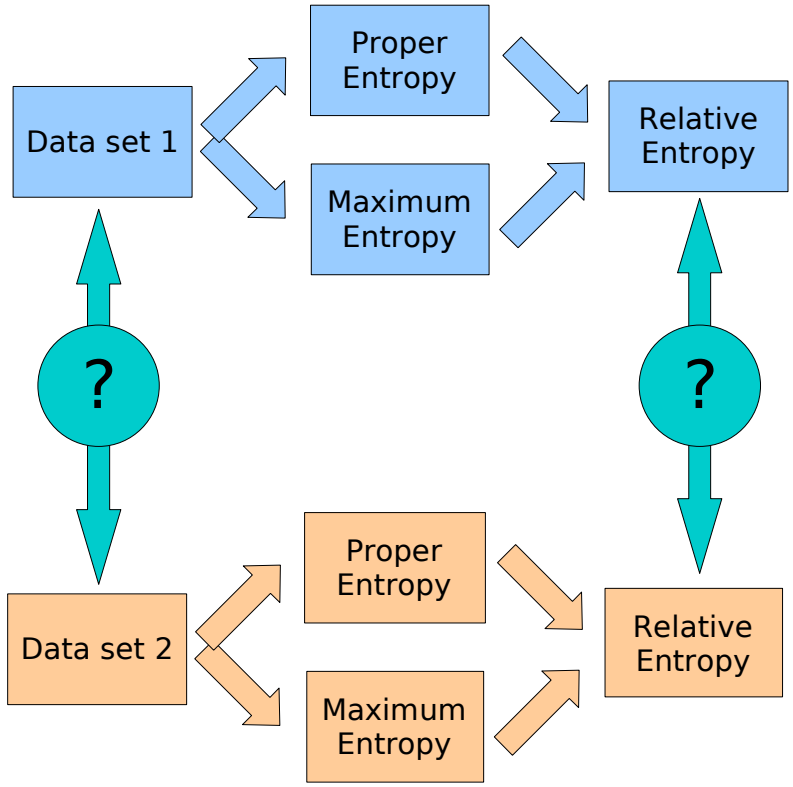
$$H_p = -\sum_{1}^{i} p(z_i) \log_2 p(z_i)$$

- **Maximum Entropy H_{max}: Maximum entropy is reached when the elements of a data-set are distributed uniformly**
- **Relative Entropy H_{rel}: Relation between proper entropy and maximum entropy:**

$$H_{rel} = \frac{H_p}{H_{max}}$$

Questions and problems regarding entropy

- **All kinds of entropies are statements over a specific data set. But how is it possible to construct relations between *different* data sets?**

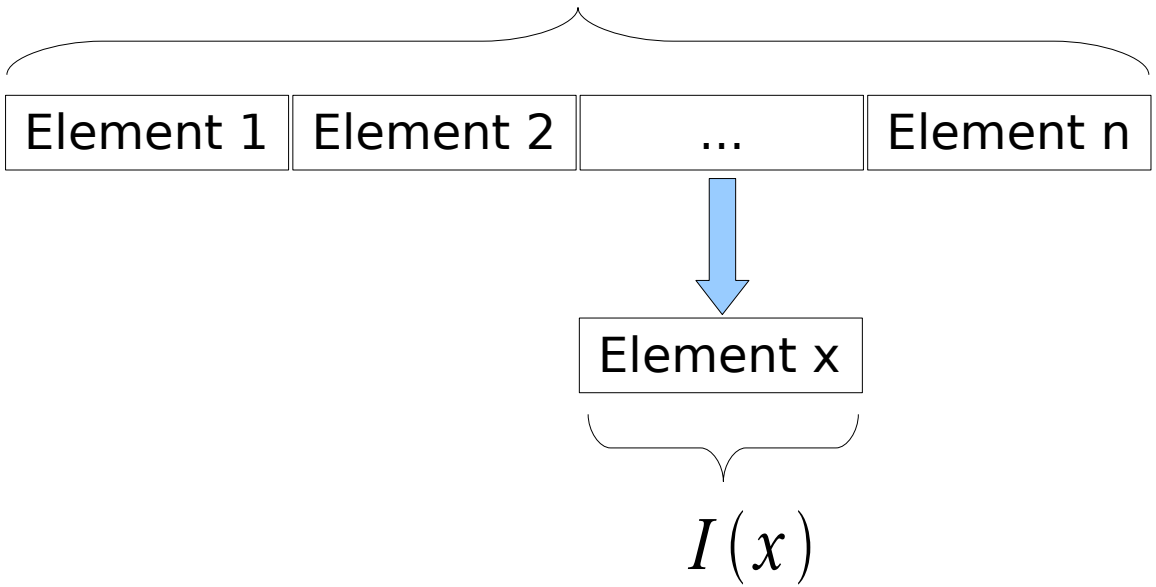


Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Extending entropy

- **Solution: partial information**
- **In Dialectometry: Information is the amount of differences between data sets, for example sites**

$$I(s) = \sum_{1}^{n} I_n$$



Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Extending entropy

- **Information for every element:**

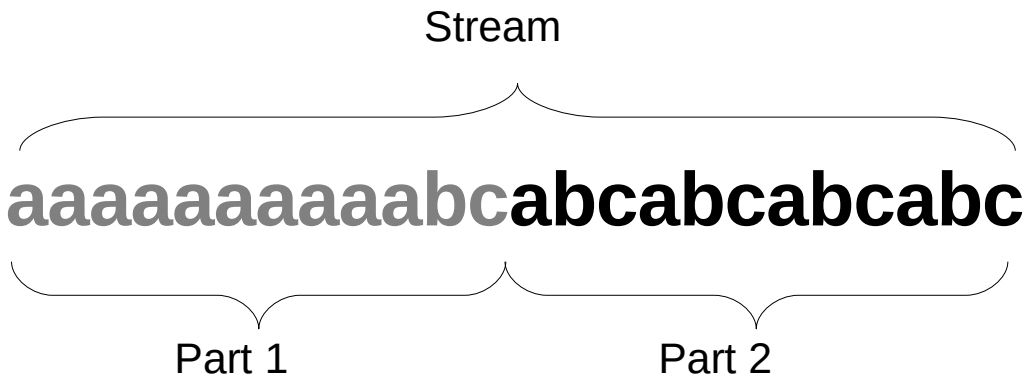
$$I(x) = \sum_{1}^{i} \log_2 p(z_i)$$

where 1 ... i iterates over *every* character of the element

- **Partial information:**

$$I_p(x, s) = \frac{I(x)}{I(s)}$$

Extending entropy



	Stream	Part 1	Part 2
Alphabet	{a, b, c}	{a, b, c}	{a, b, c}
Information	33.5	12.3	21.2
Partial Information		0,37	0,63

(All values are rounded)

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

First results: bigram-analysis

- **A vowel bigram *v?* is defined as a bigram which starts with a vowel**
- **v is element of {a, e, i, o, u}**
- **? is the phone directly after the vowel**
- **Partial information I_p(site, v) can be calculated:**

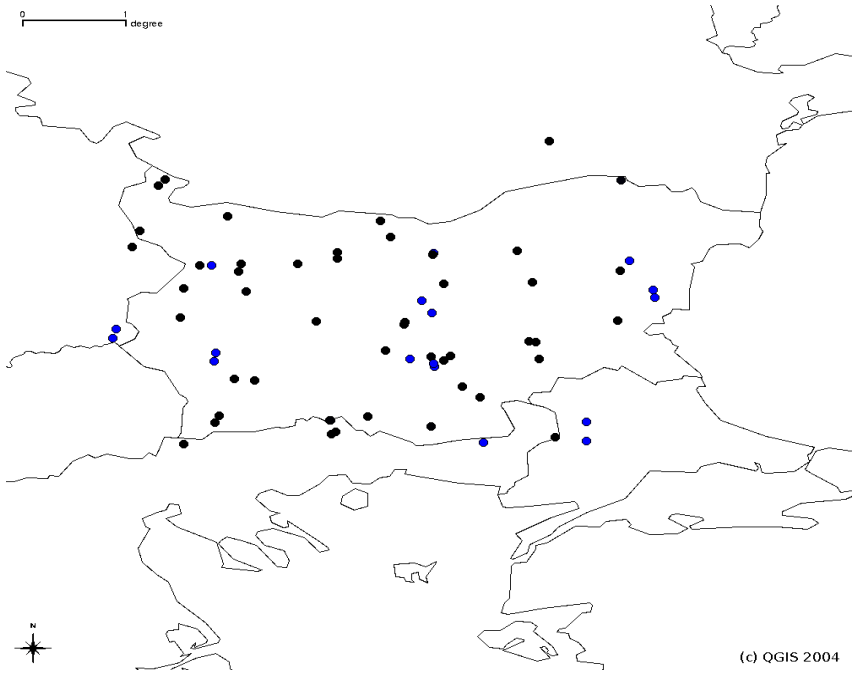
$$I_p(site, v) = \frac{I(BGs_v)}{I(AllBGs_v)}$$

where BGs_v is the bigram stream of the individual site and $AllBGs_v$ is the bigram stream of the whole data set

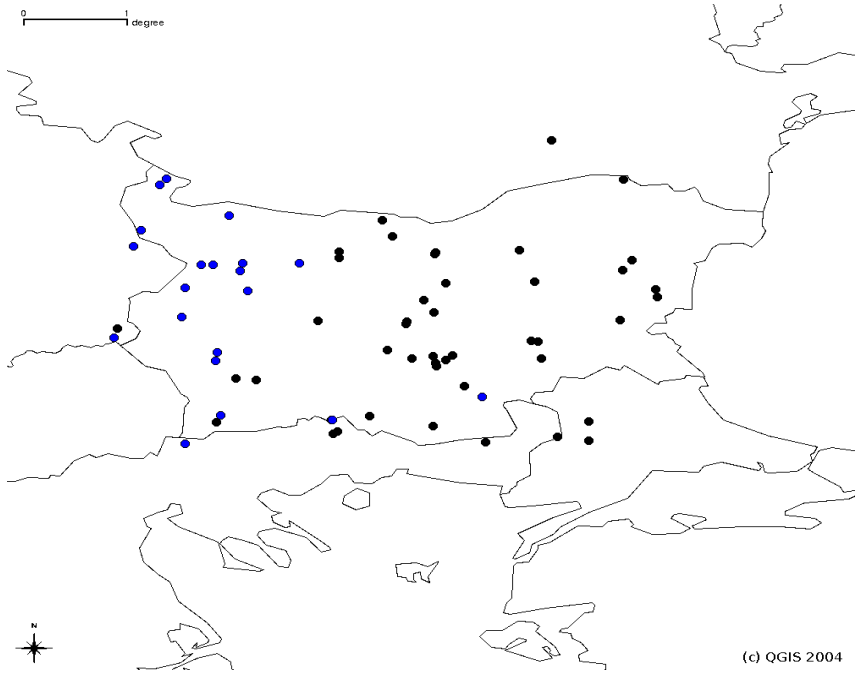
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

First results: bigram analysis

a?



e?



Appearance of bigrams:

e?	-	4697
a?	-	2869
i?	-	2096
u?	-	1930
o?	-	1919

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

First results: bigram analysis - ToDo, next steps

Bigrams:

- Better clustering: Dividing the results of the bigram analysis into delimited areas, using different colors
- Using other bigrams than the vowel bigrams
- Using trigrams instead of bigrams
- Applying the partial information to other data contexts
- Other kinds of visualization

In general:

- Using other logarithms
- Weighting of elements

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

References

- **Project-Homepage: http://www.sfs.uni-tuebingen.de/dialectometry/**
- **Wikipedia article about entropy: http://de.wikipedia.org/wiki/Entropie_%28Informationstheorie%29**
- **Lyre, Holger. Informationstheorie, München 2002**
- **Klimant, Herbert u.a. Informations- und Kodierungstheorie, Stuttgart 2003**